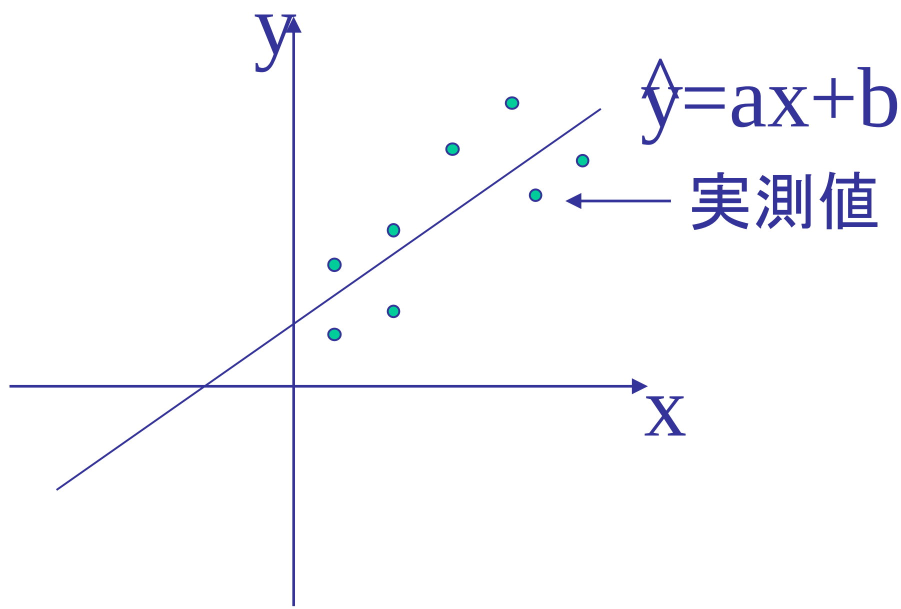
最小自乘法

回帰分析

- 説明変数と目的変数の関係のモデルを推定する統計的手法
  - ある変量(説明変数)とその変量に対する望みの結果(目的変数)の値がいくつか与えられる
- 説明変数が2つ以上ある場合には，重回帰分析と言う

回帰分析

- 説明変数(x)と目的変数(y)の関係の実測値がいくつか与えられたとする(下図)
- x に対する y の予測値 y は，y = ax + b で与えられる
- この直線の最適な係数 a, b を求めることによって，変数 x と y の関係が予測する
- 係数 a, b を回帰係数と呼ぶ



最小自乗法

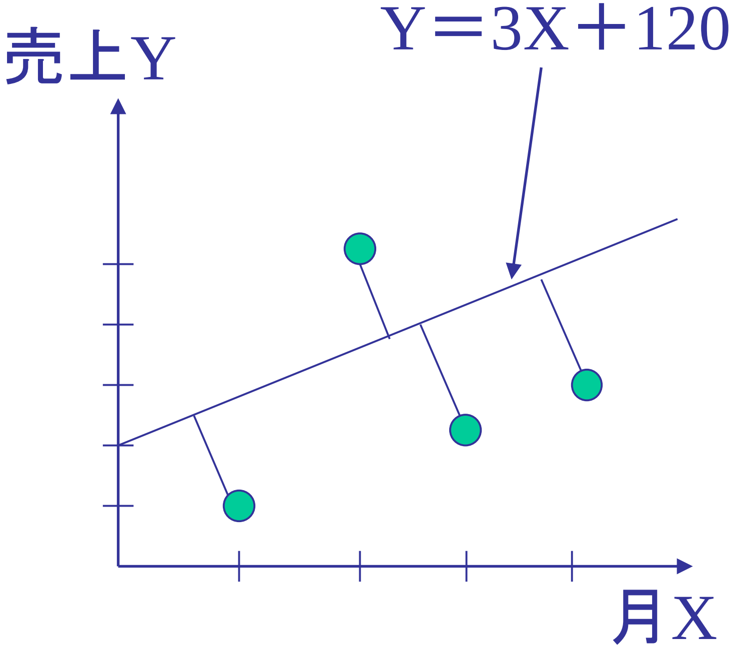
- 回帰係数を求める手法
- 実測値 y_i と予測値 $\hat{y}_i$ $(=ax_i+b)$ の差(残差)の自乗和を最小にするような回帰係数 a,b を求める
  - 実測値のデータの個数:N
  - x_i: i 番目の x の実測値 $(i = 1 \cdots N)$
  - y_i: x_i に対する y の実測値とする $(i = 1 \cdots N)$
  - a: 直線の傾き
  - b: 切片

最小自乗法

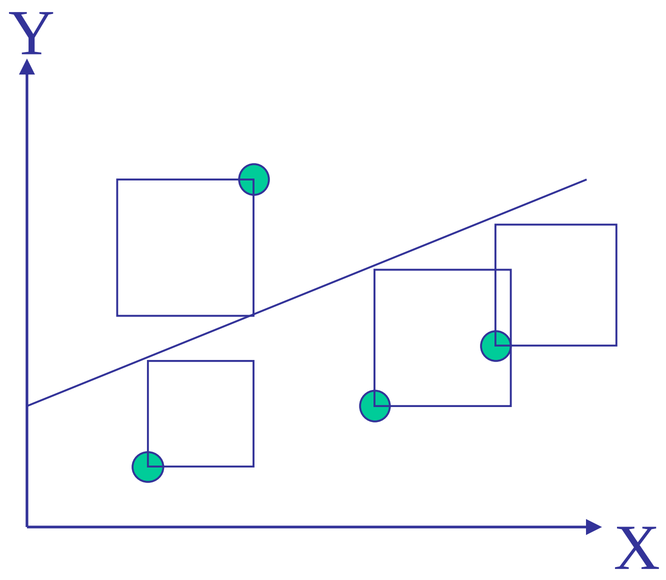
5月の売上は？

月	売上
1月	110
2月	150
3月	120
4月	130
5月	？

実測値



最小自乗法



y=a+bx+誤差

誤差の自乗和（図の正方形の和）を最小にするようにして未知係数aとbを求めるので最小自乗法と呼ぶ

最小自乗法

- 残差の自乗和: Q

$$Q = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{N} (y_i - (ax_i + b))^2 \qquad (0)$$

最小自乗法

- Qを最小とする a, b は, $\frac{\rho Q}{\rho a} = 0$ かつ $\frac{\rho Q}{\rho b} = 0$ を満たす

$$\frac{\rho Q}{\rho a} = -2\sum_{i=1}^{N} x_i(y_i - (ax_i + b)) = 0 \qquad -(1)$$

$$\frac{\rho Q}{\rho b} = -2\sum_{i=1}^{N} (y_i - (ax_i + b)) = 0 \qquad -(2)$$

(1)式を整理すると

$$\sum_{i=1}^{N} x_i(y_i - (ax_i + b)) = 0$$

$$\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} (ax_i x_i + bx_i)) = 0$$

$$\sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i x_i = b\sum_{i=1}^{N} x_i \qquad -(3)$$

最小自乗法

- 前ページ(2)式を整理すると

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} a x_i - \sum_{i=1}^{N} b = 0$$

$$\sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i - Nb = 0$$

$$\sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i = Nb$$

$$b = 1/N \left(\sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i\right) \qquad - (4)$$

(4)式を(3)式に代入して

$$\sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i x_i = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i\right)\sum_{i=1}^{N} x_i$$

$$N\sum_{i=1}^{N} x_i y_i - aN\sum_{i=1}^{N} x_i x_i = \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i$$

$$a\left(\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i - N\sum_{i=1}^{N} x_i x_i\right) = \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i - N\sum_{i=1}^{N} x_i y_i$$

$$a = \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i - N\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i - N\sum_{i=1}^{N} x_i x_i} \qquad (5)$$

以上のようにして，回帰係数を決定

この時のa,bが予測値と実測値の差の自乗を最小にするような，予測値の式の係数

```
#include<stdio.h>
main()
{
    FILE *infile;
    int i, n;
    double a, b, x, y, x_sum, y_sum, xy_sum, xx_sum;
    if ((infile=fopen("input.dat","r"))==NULL) {
        printf("can't open file ¥n");
        exit;
    }
    n=0;
    xy_sum=0;
    x_sum=0;
    y_sum=0;
    xx_sum=0;
    while( fscanf(infile, "%lf%lf¥n", &x, &y) != EOF ) {
        xy_sum=xy_sum+x*y;        /* x*y の和 */
        x_sum=x_sum+x;            /* x の和 */
        y_sum=y_sum+y;            /* y の和 */
        xx_sum=xx_sum+x*x;        /* x*x の和 */
        n=n+1;                    /* データの数 n */
    }
    a=(x_sum*y_sum-n*xy_sum) / (x_sum*x_sum-n*xx_sum);
    b=(y_sum-a*x_sum)/n;
    printf("y = ax + b¥n");
    printf("a = %lf¥nb = %lf¥n", a, b);
    fclose(infile);
}
```